Procedia Computer Science
Volume 88, 2016, Pages 125–130

7th Annual International Conference on Biologically Inspired
Cognitive Architectures, BICA 2016

Implementing a Seed Safe/Moral Motivational System with the Independent Core Observer Model (ICOM)

Mark R. Waser[1] and David J. Kelley[2]
[1]*Digital Wisdom Institute, Vienna, VA, USA*
[2]*Artificial General Intelligence Inc, Kent, WA, USA*
Mark.Waser@Wisdom.Digital, David@ArtificialGeneralIntelligenceInc.com

Abstract

Arguably, the most important questions about machine intelligences revolve around how they will decide what actions to take. If they decide to take actions which are deliberately, or even incidentally, harmful to humanity, then they would likely become an existential risk. If they were naturally inclined, or could be convinced, to help humanity, then it would likely lead to a much brighter future than would otherwise be the case. This is a true fork in the road towards humanity's future and we must ensure that we ***engineer*** a safe solution to this most critical of issues.

Keywords: Autopoiesis, Bootstrapping, Consciousness, Emotion, Enactive, Machine Intelligence, Safe AI, Self

1 Introduction

The first step towards engineering ***anything*** is to fully specify the requirements for and desired behavior of the desired system/solution. It is truly scary therefore that, for safe machine intelligences, no one has done even that – much less outlined a credible path towards getting there. This paper, therefore, will outline one possible set of such requirements and desired behaviors and, further, outline a design and implementation plan for an engineering approach that will meet those requirements and produce those behaviors. Moreover, it will do so using an approach that is inspired by and compatible with the well-explored state space of human intelligence rather than a de novo approach based upon questionable "rationality" and relying upon a perfect (and perfectly understood) world.

The most common approach to machine intelligence, probably most widely illustrated by Asimov's Three Laws of Robotics [1], is that they should fulfill the needs of and be subservient to humanity. Asimov, of course, proposed his laws because they raised such fascinating issues that they practically guaranteed a good story. On the other hand, fear of the potentially devastating effects of "UnFriendly" intelligences prompted Yudkowsky to propose [2] a novel "cleanly causal hierarchical goal structure" logically derived from a singular top-level super-goal of "Friendliness" – presumed sufficient to ensure that intelligent machines will always "want" what is best for us. Unfortunately,

Selection and peer-review under responsibility of the Scientific Programme Committee of BICA 2016

doi:10.1016/j.procs.2016.07.415

Yudkowsky not only believes that fully defining "Friendliness" is basically insoluble without already having a Friendly AI (FAI) in place but he wants and expects his first FAI to safely figure out exactly what its goal actually is -- invoking his claimed “structurally Friendly” goal system’s “ability to overcome mistakes made by programmers” and even “overcome errors in super-goal content, goal system structure and underlying philosophy.” We have previously [3] pointed out all of the problems with this approach including the facts that it has a single point of failure by requiring protection of the ***changing*** singular goal from corruption due to error or enemy action.

Further, along with Wissner-Gross[4], we strongly contend [5] that the entire concept of limiting freedom and options (to another’s desires) is inconsistent with intelligence and argue that designing the intelligence to act “morally” (rather than subserviently) is critically necessary for a stably safe solution. Numerous others have agreed but as we have noted previously [6], there is an almost total unwillingness to take on the necessary first step of defining human values or morality. Instead, while many have bemoaned the supposed “complexity and fragility” of human values [7] and argued [8] that “any claims that ethics can be reduced to a science would at best be naive” and “engineers will be quick to point out that ethics is far from science”, they then propose a seemingly endless proliferation of, what we would contend to be unrealistic, machine learning research projects for analyzing human value judgments and morality from examples – ranging from Yudkowsky’s “Coherent Extrapolated Volition” [9] to Russell’s “inverse reinforcement learning” [10].

2 Requirements & Desired Behaviors

The sole requirement of “morality” is all that is necessary to prevent the most egregious results. As long as machine intelligences follow the dictates/requirement of morality, they should not become the existential risk that so many fear. As pointed out by James Q. Wilson [11], the real questions about human behaviors are not why we are so bad but “how and why most of us, most of the time, restrain our basic appetites for food, status, and sex within legal limits, and expect others to do the same.” The fact that we are generally good even in situations where social constraints do not apply is because we have evolved to cooperate [12-15] by developing a “moral sense” that virtually all of us (except sociopaths and psychopaths) possess and are constrained by (just as we wish intelligent machines to be constrained) [16-19].

Uncaught and/or unpunished immorality frequently confers substantial advantages upon the perpetrator at a cost to others and society as a whole – the exact definition of selfishness. Social psychologist Jonathan Haidt’s definition of the function of morality [20] – to regulate or suppress selfishness and make cooperative social life possible – explains virtually every moral behavior that has evolved as well as the differences between the moral behaviors of societies living under different circumstances. We shall treat this definition as Kant’s Categorical Imperative – an action that should be universalized and taken regardless of circumstance. This meets our originally specified requirements [21] of a universal ethical system that is simple, safe, stable, self-correcting and sensitive to current human thinking, intuition and feelings.

We can, however, do much better than merely implementing the requirement/restriction of morality. Instead of merely preventing harmful actions, we should also promote beneficial ones. As morality is basically about balancing what is “best” for a given individual vs. others and society in general, we would like to extend this by actively promoting what is best for others and society wherever this does not severely conflict with an intelligence’s own self-interest. This is useful for the intelligence itself because it increases both the general advantages of society around it and the likelihood that others will specifically befriend it and assist it with its goals. In particular, it would be particularly worthwhile to build and support an open community of moral “people” that believe that helping each other is the best way to serve one’s own interests. While philosophers have long debated what we “ought” to do, simply recognizing these facts offers concrete suggestions. In addition to